Contents

List of Figures

List of Tables

Chapter 1

Introduction

Description of:

- Motivation to investigate nearby young open clusters.
- The importance of the IMF
- Current state of knowledge in dynamical simulations.
- The problematic of constrains in dynamical theories.

It must be clear which is the objective

Description of the Nearby open clusters and their properties.

1.1 The DANCe project

- List of open clusters in the DANce project.
- the importance of the pleiades, why we restrict to it.

It must be clear what are the limitations, the boundaries in which the objective will be searched

Description of the current methodologies used to address the question mentioned previously.

- The works of Sarro, Krone-Martins, Malo, Gagne etc.

-The advantages and caveats of the previous methodologies.

-It must be clear the necessity of a new perspective

The proposal we made. The use of Bayesian Hierarchical Models. Benefits and issues of BHM.

Description of the advantages of BHM.

-¿ It must be clear that BHM are the best choice.

Description of the practical issues needed to be solved in order to use BHM.

MCMC techinques and PSO.

-¿ It must be clear that MCMC methods are the best option.

Brief descriptions of our results and how they impact our current knowledge.

-¿ It must be clear that we attained the objective: The pleiades velocity, spatial and mass distributions.

Chapter 2

The Pleiades as a benchmark

2.1 Generalities

Description of the current knowledge on the pleiades cluster Age, distance, metallically reddening, number of members,

2.1.1 The distance controversy

2.2 Spatial Distribution

2.3 Velocity Distribution

2.3.1 Radial Velocity

2.4 Luminosity Distribution

2.5 Mass Distribution

2.6 The current dynamical scenario

2.7 The Pleiades DANCe DR2.

Chapter 3

Bayesian formalism

This chapter provides a general introduction to probability theory and its application to parametric inference. Since the objective of this work is to infer the probability distributions of the cluster properties (e.g. luminosity and velocity), I give reason to prove that the Bayes' theorem provides the proper framework for the inference of the parameters governing these distributions. Later in this chapter, I describe the reason for which the Bayesian Hierarchical Models are the best option to parametric inference under the Bayesian framework.

In the following Sections I will describe in detail the assumptions I made to model the data and to select the prior distributions. The two final Sections of this Chapter focus on: the practical issues related to the sampling of the cluster distributions, and the description of details and assumptions embed in the codes I developed.

3.1 Introduction to probability theory.

Uncertainty and probability are closely entangled. Anything we measure has an associated uncertainty, otherwise is not a complete measurement [1]. The term uncertainty must not be confused with the term error, which refers to the difference between the measured value of the quantity and the *true* value[2] of it [**?**]. It is commonly agreed that uncertainty of a measurement can be expressed in a

[1] Upper and lower limits are examples of incomplete measurements.

[2] The true value is that which ideally results when the uncertainty tends to zero.

probabilistic basis [**?**]. This means that whenever we measure a quantity, lets say a, the distribution of the repeated measurements of a, is a probability distribution function, $p(a)$. As any other probability distribution, $p(a)$ satisfies the following properties:

Property 1 It has units, those of the inverse of a.

Property 2 $p(a) \geq 0$.

Property 3 $1 = \int_a p(a)da$.

These properties hold regardless of the dimension of a, it means that the joint uncertainty of all measured quantities of an object is also a probability distribution. Furthermore, they also hold if the probability distribution is conditioned in any other quantity. Lets imagine that we measure the positions, projected in the plane of sky (the plane perpendicular to the line of sight), of one star, these measurements are conditioned in the magnitude (brightness) of the object we measure. If the object is too bright, like the sun, it will saturate the detector and it will render the measurement useless. On the other hand, if the object is too faint we simple will not have enough photons to measure it. So, the stellar positions in the sky, which we can call a and b because they are two dimensions, are conditioned on the magnitude, c, of the object. Therefore, $p(a, b|c)$ must also satisfy:

- It has units of $a^{-1}b^{-1}$.
- $p(a, b|c) \geq 0$.
- $1 = \int_a \int_b p(a, b|c)da \cdot db$.

The link between joint and conditioned probabilities is given by the following symmetric definition:

$$\begin{aligned} p(a, b) &= p(a|b) \cdot p(b). \\ p(a, b) &= p(b|a) \cdot p(a). \end{aligned} \tag{3.1}$$

This can be further conditioned on c to obtain:

$$\begin{aligned} p(a,b|c) &= p(a|b,c)\cdot p(b|c), \\ p(a,b|c) &= p(b|a,c)\cdot p(a|c), \end{aligned} \tag{3.2}$$

If the joint probability of a and b can be factorised, this is

$$\begin{aligned} p(a,b) &= p(a)\cdot p(b), \\ p(a,b) &= p(b)\cdot p(a), \end{aligned} \tag{3.3}$$

then a and b are say to be *independent*. An alternative option is to say that a and b are *independent*, if the conditional probability of a on b is $p(a|b) = p(a)$.

The most important thing we can do with probability distributions is to integrate them. **Property 3** establish that the amount[3] of probability $p(a)$ spread over the volume of the support of a adds to one. This Property allows us to *marginalise* any non-desired variable. Lets imagine again that a and b are the measured positions of some star and we have several measurements of these positions. Then we will have the joint probability distribution of them, $p(a,b)$ (must likely it will be a bivariate gaussian but that does not matter now). If we are interested lets say in the mean value of a, we first must get rid of b. For it, we *marginalise* out b in the following way,

$$p(a) = \int_b p(a,b)\cdot db. \tag{3.4}$$

Then we compute the *expected value* of a, $E(a)$, which is identified with the mean of a once we have drawn many realisations from its probability distribution. To compute it, we add all the possible values of a weighted by their probability. This is,

$$E(a) = \int_a a\cdot p(a)\cdot da. \tag{3.5}$$

[3]Which could be infinite, like in Dirac's delta.

Once again, these last two equations (3.4 and 3.5) hold in case they are conditioned in any other measurement. For example, the magnitude of the object, as in our previous analogy. Notice however that, once the brightness of the object lay within in the dynamic range of the detector, a and b became *independent* of the magnitude.

It is important to recall that the term measurement, and its unavoidable uncertainty, refer not just to directly measured quantities, like the photons (counts) and pixels in a CCD, but also to indirect measurements. Stellar magnitudes and positions in the sky, for example, are indirect measurements derived from the direct measurement of photons, pixels and telescope arrangement. This generalisation also applies to the measurement of parameters in any physical or statistical model, like the one I will describe in the following Section.

3.1.1 Bayes theorem

The definition of conditioned probability (Eq. 3.2) leads to the Bayes' theorem:

$$p(a|b,c) = \frac{p(b|a,c)\cdot p(a|c)}{p(b|c)}. \tag{3.6}$$

Integrating on a we find that,

$$\begin{aligned} p(b|c)\cdot\int_a p(a|b,c)\cdot da &= \int_a p(b|a,c)\cdot p(a|c)\cdot da \\ Z \equiv p(b|c) &= \int_a p(b|a,c)\cdot p(a|c)\cdot da. \end{aligned} \tag{3.7}$$

In this last equation Z refers to what is known as the *evidence*. I will come back to this *evidence* in a few paragraphs. This last Eq. also illustrates that $p(b|c)$ is a normalisation constant which can be evaluated once $p(b|a,c)$ and $p(a|c)$ are known. These two terms are commonly referred as the *likelihood* ($p(b|a,c)$), and the *prior* ($p(a|c)$). Also the term the term $p(a|b,c)$ is called the *posterior*. These names arise in the context of parametric inference, as I will describe in the next paragraph. However, it worths mention that, although formally the likelihood and the prior are probability distributions in b and a, respectively, for $p(a|b,c)$ to be a probability distribution on a, it only suffices that the product of the likelihood times the prior

does not vanish everywhere or be negative anywhere[4]. In this case, they are called *improper* priors or likelihoods. If their product vanishes everywhere, which may be the case if the prior is terribly specified or if the likelihood does not take proper account of extreme data, then the posterior is not a probability distribution due to a division by zero. In any case, it makes no sense try to estimate the parameters of a model with zero evidence.

3.1.1.1 Models and parametric inference

In a broad sense, models are representation or abstraction of the knowledge about something. Sometimes the knowledge is shared by others, some time it is not. They are everywhere in our daily life: from the words we spoke every day, to the evolution of the species and the general relativity; from a kid's draw to the cosmological models. In science, however, we restrict the concept of model to a mathematical representation of the relations among the variables. If the model is parametric, the variables include the data $\mathbf{D}$, which the model attempts to model, and the parameters θ. Parameters are free variables that allow the model to describe the data. Thus a parametric model $\mathcal{M}$, can be represented as:

$$\mathcal{M} = \{f(\mathbf{D}|\theta), \quad \theta \in \Theta\}, \tag{3.8}$$

where $f(|)$ is the function that relates data and parameters, and $\theta \subset \mathbb{R}^k$ with k the dimension of θ. Parametric inference is then the act of finding the distribution of θ given the data $\mathbf{D}$.

The Bayes' theorem allows to perform parametric inference, which is called then Bayesian inference. In this context the Bayes' theorem is:

$$p(\theta|\mathbf{D}, M) = \frac{p(\mathbf{D}|\theta, M) \cdot p(\theta|M)}{p(\mathbf{D}|M)}. \tag{3.9}$$

where $\theta, \mathbf{D}$ and M are correspond, respectively, to the parameters in the model, the data which the model tries to describe, and the prior information used in the construction of the model. Whenever we have a model, we have prior knowledge over it. Actually, it can be classified in two kinds of prior information. One

[4] Although negative probabilities may have sense in quantum mechanics. See for example [**?**]

refers to the prior information conveyed in the model, which I call M. This is the information that the creator of the model uses to establish the relations among the elements of the model: variables. The second kind of prior, $p(\theta|M)$ refers to the statement the user of the model made of his/her believes about the probability distribution of the parameter values. This is indeed subjective. However, it is, in my opinion, less subjective than the former, M, prior information. At least in this last kind, the subjectivity is expressed objectively in a probabilistic, and therefore measurable way.

The likelihood of the data $p(\mathbf{D}|\theta, M)$, is a probability distribution on the data, $\mathbf{D}$. However, it is a function on the parameters, θ, which corresponds to the function f of Eq. 3.8. If we assume that data is independent from each other, which means that the probability of measuring the value of one datum is independent of the measured value of another datum, then, the joint probability $p(\mathbf{D}|\theta, M)$ can be expressed as,

$$p(\mathbf{D}|\theta, M) = \prod_{n=1}^{N} p(d_n|\theta, M), \quad n = \{1, 2, ..., N\} \tag{3.10}$$

The term $p(d_n|\theta, M)$ is the likelihood of datum d_n. This term also as the *generative* model of the data since it contains the necessary information to generate the data[5].

I interpret the Bayes' theorem, as the probabilistic way to update knowledge. It is the way, to update knowledge once we recognise the uncertainty associated to it. In my perspective, knowledge is always uncertain, even if its uncertainty is negligible given the current evidence that supports it. The Bayes' theorem helps us to update our prior believes by means of the data, once we multiply them by the likelihood of it. Then, the posterior probabilities of the parameters given the current data, became the new prior believes once more data is available. Furthermore, the Bayes' theorem also provides the objective way to compare two models and update the prior information, M, used to construct them. This is called model selection.

[5]Actually the *true* data. To generate the observed data the noise process must be also specified.

3.1.2 Model Selection

Whenever we have a data set and two or more models that attempt to describe it, the most straightforward thing to do is to compare these models. Almost always, we want to select the *best* model. Obviously the term *best* depends on the objective of research. For example, lets imagine that our data set consists of a set of measurements of the positions of an object as it moves in the sky. If we were interested in reproducing exactly the same points in the data set, the *best* model will be a polynomial with degree equal to the number of points. This polynomial will pass trough all the points.

Once we recognise the unavoidable uncertainty of the data, we realise that an exact representation of the data is of poor use. In general, we are interested in the predictive capabilities of a model, this is its ability to predict future observations rather than to replicate the ones we have. Thus, an exact representation of the observed data (an over-fitted model), will poorly describe any new data set. In this sense, an over-fitted model *memorises* the data rather than *learns* from them.

A model that *learns* from the data is that which obtains the *true* underlying relation embedded in data. This *true* underlying relation produces de data, once it is convolved (added) with the source of uncertainty. Thus, we call *deconvolution* the the process by which the *true* underlying relation is obtained. Nevertheless, we still need to select among learning models.

We can draw some help from the commonly known Ockham's razor or principle [6]. It says:

Among competing hypotheses, the one with the fewest assumptions should be selected.

Here, I identify hypotheses with models. Thus, this principle tells us we should choose the model with the fewest assumptions. I classify the assumptions of a model in two groups: fixed and free ones. The fixed assumptions belong to what I previously described as the prior information, M, used to construct the model. They render the model more physically or statistically interpretable, give it sense within a corpus of hypotheses. Free assumptions correspond to the parameters of

[6]The origin of this motto and its exact phrasing is beyond the scope of this work. I just mention that paradoxically, an ancient formulation is attributed to Ptolomey: "We consider it a good principle to explain the phenomena by the simplest hypothesis possible" [**?**]

the model. They give it more flexibility to fit the data. For example, in the case of a straight line model, a fixed assumption is that the data is linearly related, whereas the free assumptions correspond to the slope and ordinate at the origin. A linear model and a quadratic model in which the constant term has been fixed, have the same number of free parameters (assumptions) but clearly the second one has an extra fixed assumption. Therefore, choosing the model with fewer free parameters does not necessarily means choosing the model with the fewest assumptions.

One of the great advantages of the Bayesian methodology is that it incorporates directly Ockham's principle. Suppose we want to compare two models M_1 and M_2, which we assume describe the data set $\mathbf{D}$. Each model has prior probabilities, $p(M_k)$ and likelihoods $p(\mathbf{D}|M_k)$ (with $k = 1, 2$). The prior probabilities reflect our believes about the fixed assumptions mentioned before. On the other hand, the likelihood of the data, given the model, is related parameters (the free assumptions) and priors within a model. It corresponds to the *evidence* of the model, Eq. 3.7. This evidence in terms of the model parameters, θ_k, is

$$p(\mathbf{D}|M_k) = \int_{\theta_k} p(\mathbf{D}|\theta_k, M_K) \cdot p(\theta_k|M_k) \cdot d\theta_k. \tag{3.11}$$

The Bayes' theorem applied to models instead of individual parameters tells us that

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k) \cdot p(M_k)}{p(\mathbf{D})}. \tag{3.12}$$

with $k = 1, 2$. Since there are only two models, their prior probabilities are related by $p(M_1) = 1 - p(M_2)$. Therefore,

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k) \cdot p(M_k)}{p(\mathbf{D}|M_1) \cdot p(M_1) + p(\mathbf{D}|M_2) \cdot p(M_2)}. \tag{3.13}$$

From this last Equation, the ratio of the posterior distributions is:

$$\frac{p(M_1|\mathbf{D})}{p(M_2|\mathbf{D})} = \frac{p(\mathbf{D}|M_1) \cdot p(M_1)}{p(\mathbf{D}|M_2) \cdot p(M_2)}. \tag{3.14}$$

This ratio provides an objective measure of how better model M_1 is when compared to model M_2, under the measure provided by the data $\mathbf{D}$ by means of the evidence. When both prior probabilities $p(M_1)$ and $p(M_2)$ are set equal, the ratio of posteriors

equal the ratio of likelihoods. This is known as the *Bayes factor*[7]. Even in the equal priors case, the evidences themselves, Eq. 3.11, embody the Ockham's principle. Indeed, each evidence is a measure of the prior times the likelihood, this time for parameters in a single model. The larger the number of parameters (assumptions), the larger the volume, in parametric space, over which the likelihood of the data spreads. Since the likelihood is not a probability distribution on the parameters, it does not integrate to one, even if the priors are uniform. The evidence also penalises the assumptions made in the priors of the parameters. The most concentrated the prior is the less of the likelihood contributes to the evidence.

Thus, the Bayes' theorem is the way to update knowledge, either if it refers to models or to parameters within a model.

3.1.3 Membership probability

In the previous Section, I derived, by means of the Bayes' theorem, the probability of models M_1 and M_2 given the data $\mathbf{D}$. Now, I describe the same problem but instead of the likelihood of a data set I do it for a single datum. This is, the probability of model M_1 or M_2, given the datum $\mathbf{d}$. This is known as the membership probability of the datum $\mathbf{d}$ to model or class, M_k ($k = 1, 2$). The Bayes' theorem in this case is,

$$p(M_k|\mathbf{d}) = \frac{p(\mathbf{d}|M_k) \cdot p(M_k)}{\sum_{k=1}^{2} p(\mathbf{d}|M_k) \cdot p(M_k)} \tag{3.15}$$

3.2 Bayesian Hierarchical Models

3.2.1 Generalities

Comparisons with other techniques in IA. It must be clear the BHM are the only option.

3.2.2 Examples

Its applications in IA. Its applications in astrophysics.

[7]For a similar derivation and some example see [**?**]

3.2.3 Graphical representation:Probabilistic Graphical Models

3.3 Modelling the data

Positions, proper motions, photometry

3.3.1 Missing values

3.3.1.1 Missing value pattern

3.3.2 The field population

3.3.3 The cluster population

3.4 Priors

3.5 Sampling the posterior distribution

Description of the techniques used to obtain samples from the posterior distribution.

History and used versions.

3.5.1 PSO

3.5.1.1 The charged PSO

3.5.2 MCMC

3.5.2.1 Generalities

3.5.2.2 Flavours

HMC,NUTS,Gibbs, Metropolis-Hasting, Affine invariant, stretch-move, MultiNest. It must be clear why we choose emcee and multinest

3.5.2.3 Convergence

3.5.2.4 The evidence

3.6 Codes

3.6.0.5 The modified charged PSO

3.6.0.6 Improvements of emcee

3.6.0.7 The GMM with missing values

3.6.1 Parallel implementations

Description of the implementation. MPI, python stan, etc. explain in detail the difficulties faced at implementing the different codes in the different servers.

Chapter 4

Results

4.1 Performance of the classifier

Compare and explain the differences with Stauffer members.

4.2 Velocity distribution

4.3 Spatial distribution

4.4 Luminosity distribution

4.5 Mass distribution

4.5.1 The mass-luminosity relation

4.6 The mass distribution on time

4.7 The phase space

4.8 Updating the previous knowledge

Chapter 5

Conclusions and Future Work

Appendix A

Appendix Chapter Title

In the following we include ....